MAY 15, 2021

EARLY DETECTION OF CERVICAL CANCER BASED ON BEHAVIOUR

ASSIGNMENT 2

We certify that this is all our own original work. If we took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in our submission. We will show we agree to this honor code by typing "Yes": Yes

SANJAY SAJEESH KRISHNAN
S3776159

Table of Contents:

Abstract

The aim of this report is to determine the behavioural indicators of an individual that might lead to the individual having Cervical Cancer. This dataset is based on an analysis partaken by Sobar, Machmund and Wijaya for the STIKES institute in Indonesia. During my analysis of this particular dataset, I first cleaned the data, then explored the relationships between the behavioural indicators and then performed Machine Learning Techniques such as KNN Classification and Decision Tree Classification. The accuracy and errors of each model was taken into consideration and upon further analysis, the best model was found for this dataset.

Introduction

Cervical Cancer is the fourth most common cancer prevalent in women. Fortunately, Cervical Cancer is one of the most treatable forms of cancer if it is detected early. Even if the cancer is detected in the latter stages, it can still be controlled to a certain extent. There are many causes of Cervical Cancer, but most cases are due to a virus called HPV (High-Risk Human Papillomaviruses). This report consists of two major sections, the Data Exploration section and the Data Modelling section which is explained quite extensively with the aid of two distinct classification models. Finally, this report also further showcases the behavioural indicators that may result in the individual having cancer in the data exploration section and with great precision.

Dataset Description and Methodology

The Cervical Cancer dataset used in this project is extracted from the UCI Machine Learning Repository (Sobar, Machmund & Wijaya). It contains the data collected from 72 individuals and has 19 descriptive features and one target feature (ca_cervix). Upon analysing the raw data, there were no traces of Null values or whitespaces found. After satisfying all the data cleaning checklists, I did the analysis of the data provided. To analyse the data, I used the inbuilt libraries such as Pandas, NumPy, Seaborn and ScikitLearn.

Data Exploration

Univariate Exploration

Adhering to the specifications, I have analysed 15 attributes out of the possible 20 and will explain and expand on a selected 10 of them.

Behaviour SexualRisk Column

The behaviour_sexualRisk column has a minimum value of 2 and a maximum value of 10. As observed in the provided bar graph, a majority of individuals part of this census have a behaviour_eating value of 10 and the least value (excluding zero values) of 2 & 6.

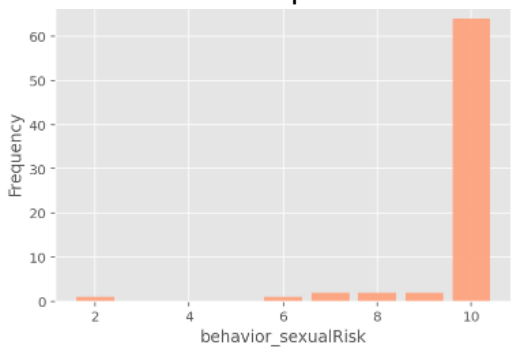


Figure 2.1.1: Barchart of the frequencies of behaviour_SexualRisk

Behaviour Eating Column

The behaviour_eating column has a minimum value of 3 and a maximum value of 15. As observed in the provided bar graph, a majority of individuals part of this census have a behaviour_eating value of 15 and the least value (excluding zero values) of 3.

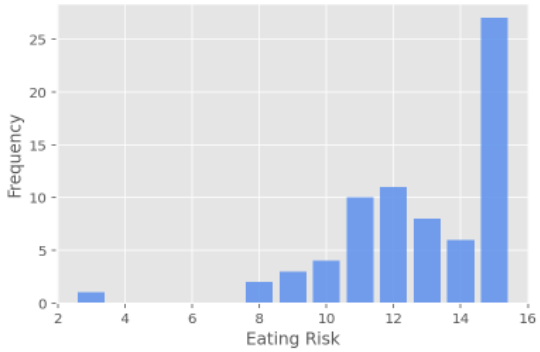


Figure 2.1.2: Barchart of the frequencies of behavior_eating

Intention_Aggregation Column

We can analyse the intention_aggregation column with the aid a density plot. The density plot peaks at a value of 10 which indicates that most individuals have an intention_aggregation value of 10. We can

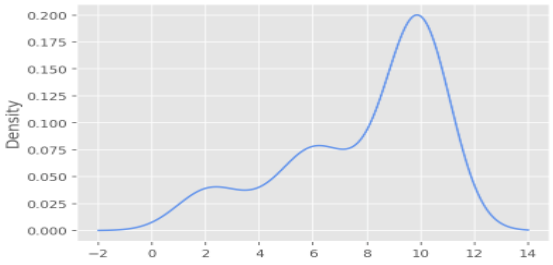


Figure 2.1.4: Density Plot of the frequencies of intention_aggregation

also identify the max value which is 10 and the min value of 2 from the graph as the curve tends to dip to zero at that point.

Intention_Commitment Column

We can analyse the intention_commitment column with the help of a density plot. The density plot peaks at 15 which indicates that a majority of individuals have a intention_commitment value of 15. We can also find the max and min value of 15 and 6 respectively from a dip in the graph towards zero.

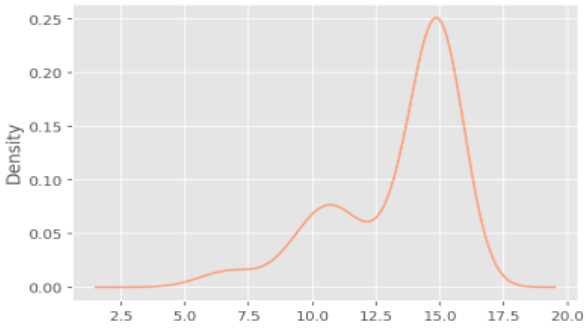


Figure 2.1.5: Density Plot of the frequencies of intention_commitment

Attitude Spontaneity Column

The attitude_spontaneity column has a minimum value of 4 and a maximum value of 10 followed by 8. As observed in the provided bar graph, a majority of individuals part of this census have a behaviour_eating value of 10 and the least value (excluding zero values) of 4 & 5.

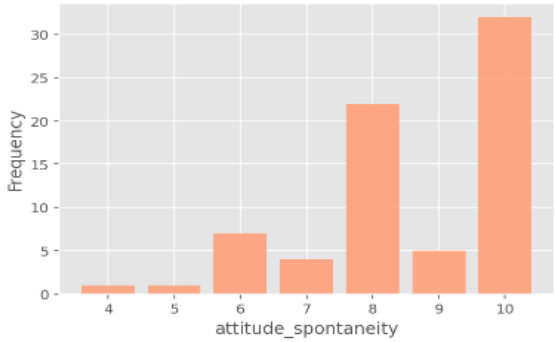


Figure 2.1.7: Barchart of the frequencies of attitude_spontaneity

Norm_SignificantPerson Column

The norm_significantPerson column has a minimum value of 1 and a maximum value of 5. As indicated by the barplot, a majority of individuals have a norm_significantPerson of 5 closely followed by a value of 1. This is a particularly prominent column as both extremes of values have a fairly similar trend and is going to aid us in the next section.

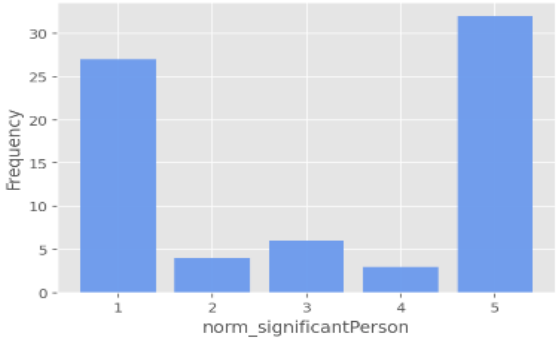


Figure 2.1.8: Barchart of the frequencies of norm_significantPerson

Motivation_Strength Column

The motivation_strength column peaks at 15 as indicated by the histogram and this indicates that a majority of individuals have a high motivation_strength. This column has a minimum value of 3 as indicated by the graph.

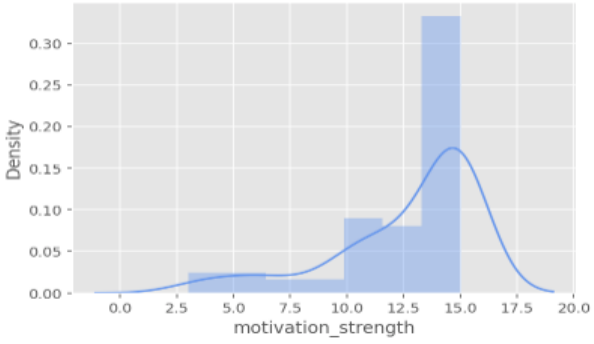


Figure 2.1.10: Histogram of the frequencies of motivation_strength

Empowerment_Knowledge Column

The empowerment_knowledge column is explored using the following bar graph. As shown in this graph, a major section of those under this census have an empowerment_knowledge of 15 followed by 13 in second. The maxima and minima as indicated by the graph is 15 and 3 respectively.

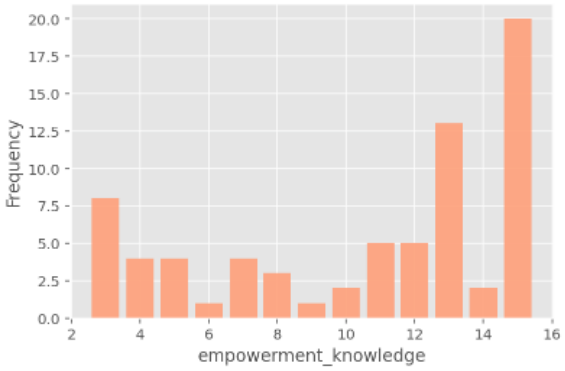


Figure 2.1.13: Barchart of the frequencies of empowerment_knowledge

Empowerment Desires Column

The empowerment_desires column is explored using the following bar graph. As shown in this graph, a major section of those under this census have an empowerment_knowledge of 15 followed by 3 in second. The maxima and minima as indicated by the graph is 15 and 3 respectively.

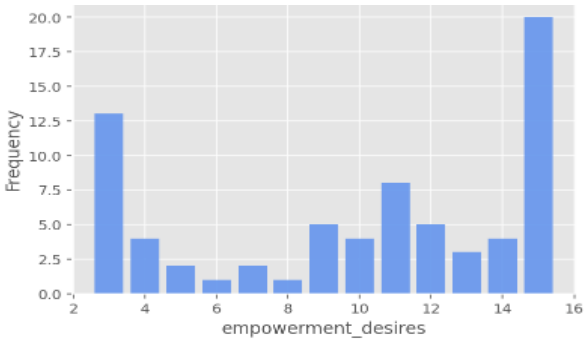


Figure 2.1.14: Barchart of the frequencies of empowerment_desires

Ca_Cervix Column

The ca_cervix column is our target value for this entire analysis. This column has two values: - 0 for non-cancerous and 1 for cancerous. Using this pie chart, we find out that 71% of individuals part of this census do not have Cervical Cancer and only 29% have Cervical Cancer.

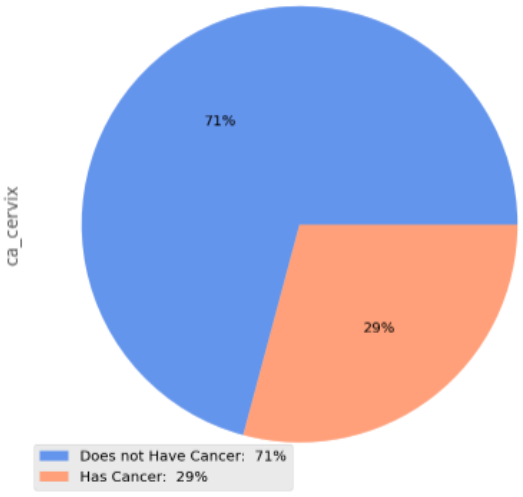
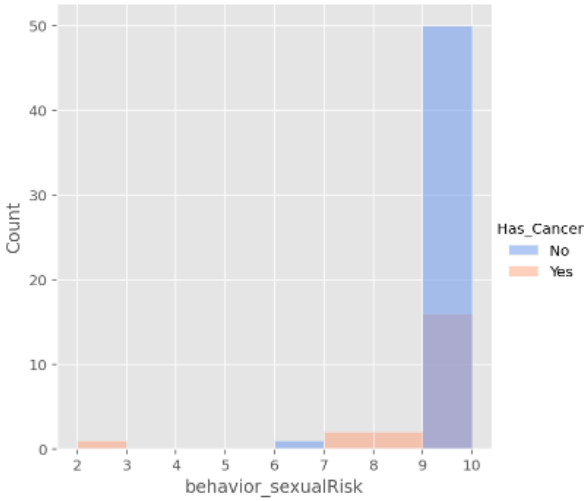


Figure 2.1.15: Piechart of the frequencies of ca_cervix

Multivariate Exploration

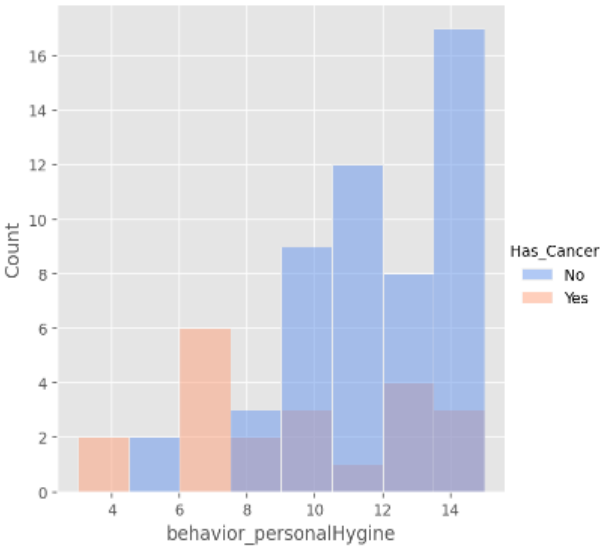
Hypothesis 1: Increased Sexual Risk increases the chance of Cervical Cancer

Findings: According to the data collected and analysed, although there seems to be an increase in number of individuals with cervical cancer with respect to the rise in SexualRisk. This hypothesis is refuted by the graph which shows that even at the peak of SexualRisk (15) there are more people without cancer than with it.
Result: Hypothesis Failed



Hypothesis 2: Increased Personal Hygiene reduces the chance of Cervical Cancer

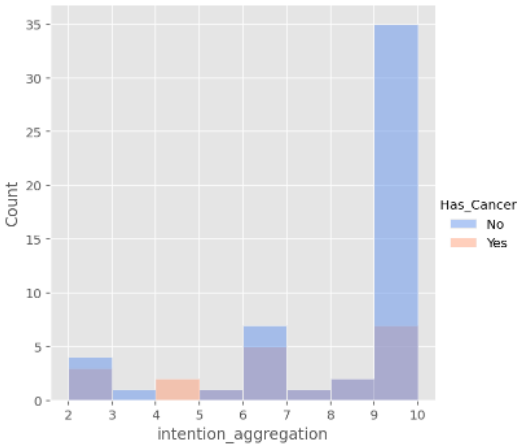
Findings: According to the data collected and processed, we come to a conclusion that an increase in personal hygiene does reduce the chance of cervical cancer. This is further proven with the help of this stacked histogram. If we are to look at the values on the extremities, we can see that at the maxima of personal hygiene, more individuals do not have cancer compared to those that do. Similarly, at the minima of personal hygiene, more people have cancer than those without.



Result: Hypothesis Passed

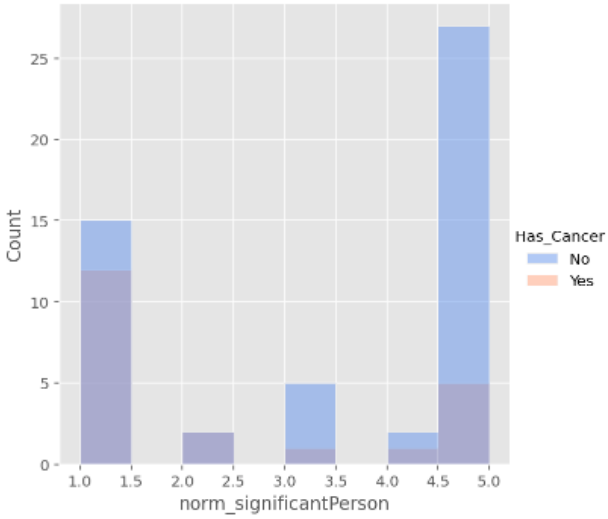
Hypothesis 3: Increased Intention Aggregation reduces the chance of Cervical Cancer

Findings: According to the data procured and upon further analysis it is found that with an increase in intention_aggregation, the chance of the individual having cancer reduces. This analysis is backed by the following histogram at its extremities. For example, at an intention aggregation of 2 the number of individuals without cancer is almost similar to those with whereas at a higher intention aggregation such as 10, we can see a substantial difference between individuals with cancer and ones without.



Result: Hypothesis Passed

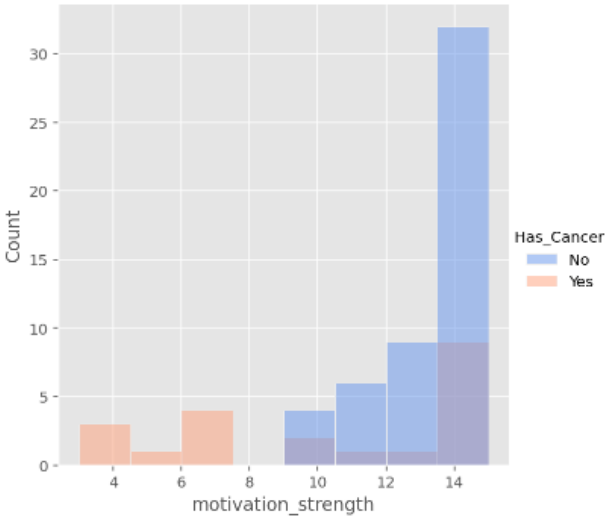
Hypothesis 4: Increased Norm Significant Person reduces the chance of Cervical Cancer



Findings: According to the data procured and upon further analysis it is found that with an increase in norm_significant_person, the chance of the individual having cancer reduces. This analysis is backed by the following histogram at its extremities. For example, at a norm_significant_person of 1 the number of individuals without cancer is almost similar to those with cancer whereas at a higher norm significant person such as at 5, we can see a substantial difference between individuals with cancer and ones without it.

Result: Hypothesis Passed

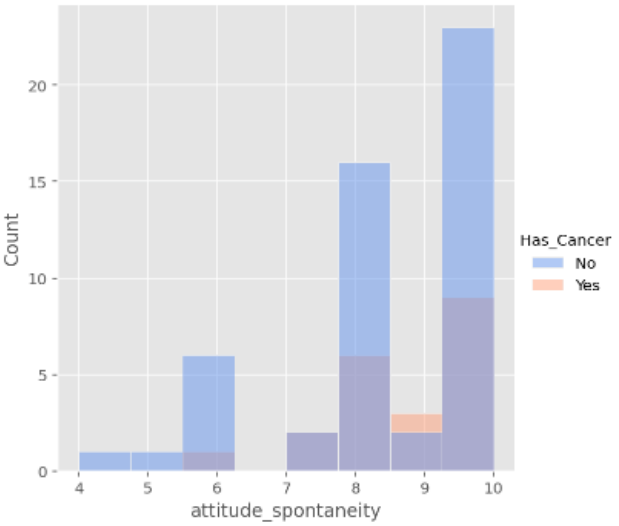
Hypothesis 5: Increased Motivation Strength reduces the chance of Cervical Cancer



Findings: According to the data obtained and upon further analysis it is found that with an increase in Motivation Strength, the chance of the individual having cancer reduces. This is backed by the following histogram. If we look at a Motivation_Strength value of 1 Motivation Strength greater than 10, we can see that the number of people without cancer surpass than those without, hence proving the hypothesis.

Result: Hypothesis Passed

Hypothesis 6: Increased Attitude Spontaneity reduces the chance of Cervical Cancer

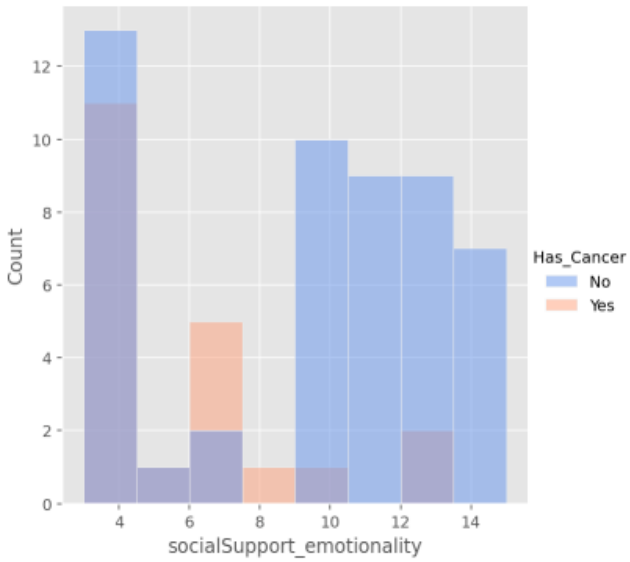


Findings: According to the data procured and upon further analysis it is found that with an increase in attitude_spontaneity the chance of the individual having cancer reduces. This analysis is backed by the following histogram. If we are to look at an attitude_spontaneity value of 10, we can see that a majority of people do not have cancer and this trend is witnessed across the entire graph where there is an increase in number of non-cancerous individuals apart from a trend change in the value of 7.

Result: Hypothesis Passed

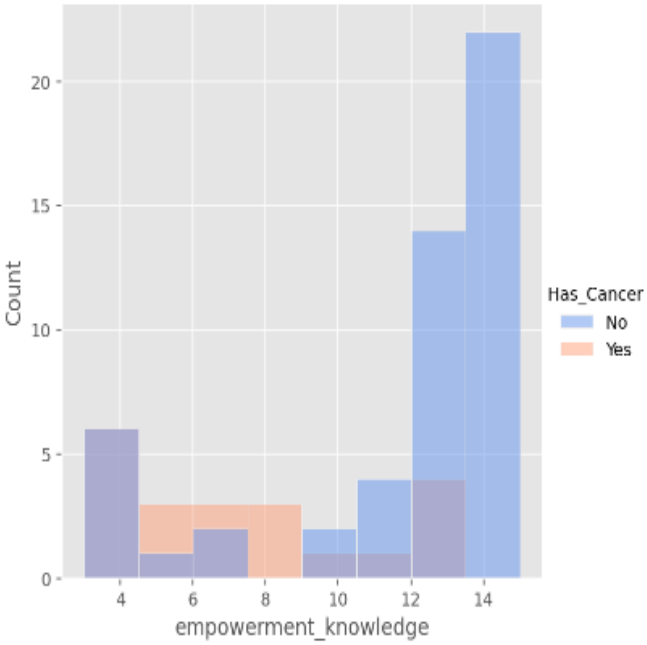
Hypothesis 7: Increased Emotional Support reduces the chance of Cervical Cancer

Findings: According to the data procured and upon further analysis is found that with an increase in Emotional Support, the chance of the individual having cancer reduces. This analysis is backed by the following histogram at its extremities. For example, at an Emotional Support of 2 the number of individuals with cancer is almost similar to those without it. Whereas at a higher Emotional Support such as 14, almost nobody has cancer, and this trend is observed throughout the graph as indicated by the increase is number of non-cancerous individuals with respect to the increase in Emotional Support.
Result: Hypothesis Passed



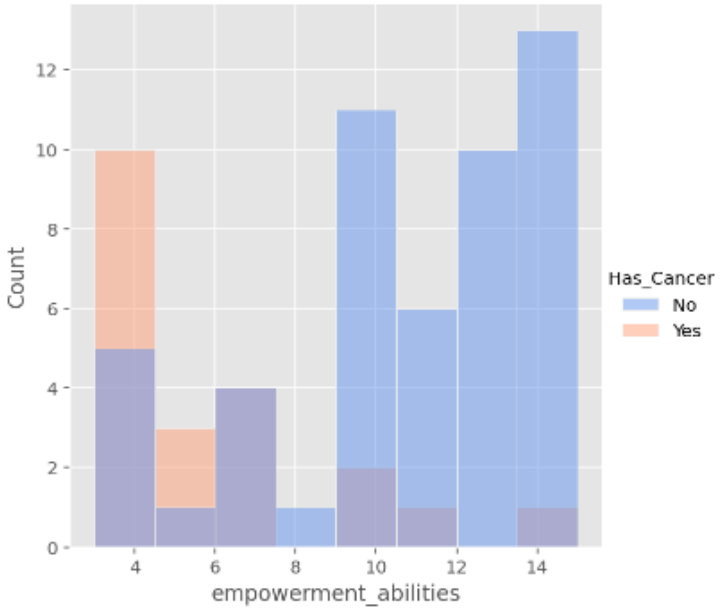
Hypothesis 8: Increased Empowerment Knowledge reduces the chance of Cervical Cancer

Findings: According to the data procured and upon further analysis is found that with an increase in Empowerment Knowledge, the chance of the individual having cancer reduces. This analysis is backed by the following histogram at its extremities. For example, at an Empowerment Knowledge of 4 the number of individuals with cancer is almost similar to those without it. Whereas at a higher Empowerment Knowledge such as 14, almost nobody has cancer, and this trend is observed throughout the graph as indicated by the increase is number of non-cancerous individuals with respect to the increase in Empowerment Knowledge.
Result: Hypothesis Passed



Hypothesis 9: Increased Empowerment Ability reduces the chance of Cervical Cancer

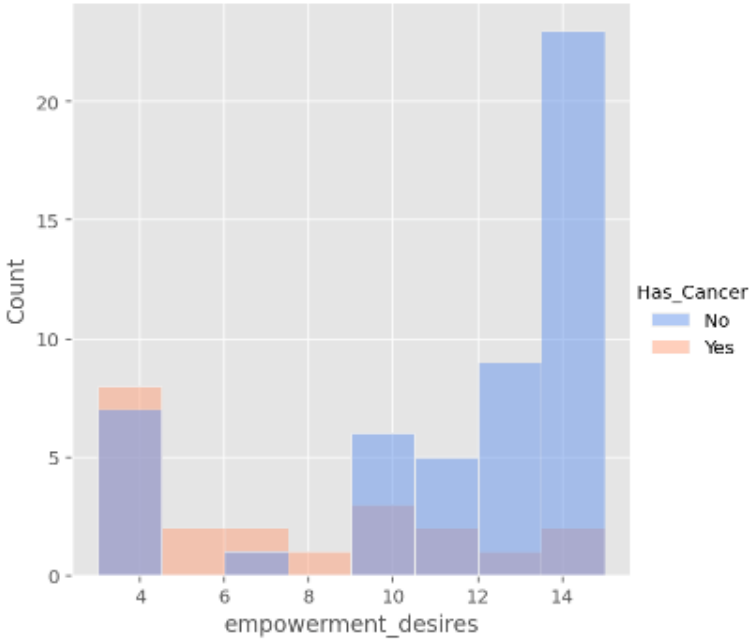
Findings: According to the data procured and upon further analysis is found that with an increase in Empowerment Ability, the chance of the individual having cancer reduces. This analysis is backed by the following histogram. For example, at an Empowerment Ability of 4 the number of individuals with cancer is significantly greater than those without. This trend is dissipated as the Empowerment Ability increases. If we are to look at the values of 10 and above, we can see a substantial difference between individuals with and those without cancer and this trend proves our hypothesis.



Result: Hypothesis Passed

Hypothesis 10: Increased Empowerment Desires reduces the chance of Cervical Cancer

Findings: According to the data procured and upon further analysis is found that with an increase in Empowerment Desires, the chance of the individual having cancer reduces. This analysis is backed by the following histogram. For example, at an Empowerment Desires of 4 the number of individuals with cancer is greater than those without. This trend is dissipated as the Empowerment Desires increases. If we are to look at the values of 10 and above, we can see a substantial difference between individuals with and those without cancer and this trend proves our hypothesis.



Result: Hypothesis Passed

Data Modelling

KNN Classification

The first modelling technique utilised in this project is the K-Nearest Neighbour classification. This classification is split into three distinct test/training splits which help us determine the precision by which our model functions. The vital K value could be any integer from $1 \leq K \leq 15$ but for my model, I chose the K value as **8**. There were multiple values greater than this that I could have chosen but doing so does not affect the accuracy of the model substantially and does lead to a significant underfitting of the model. These K value estimates have been graphed and is shown in the appendix.

KNN Split Accuracies and their Confusion Matrix

Table 1 showcases the confusion matrix and error rates of all the splits done during KNN Classification. To understand the splits, the value on the left of the '/' indicates the training set split and the one on the right represents the testing set split.

	80/20 Split	60/40 Split	50/50 Split
Confusion Matrix	[[9 1] [2 3]]	[[23 0] [2 4]]	[[28 1] [2 5]]
Error Rate	20.0 %	6.896551724137931 %	8.333333333333332 %
F1-Score for Individuals Without Cancer	0.86	0.96	0.95
F1-Score for those with Cancer	0.67	0.80	0.77

Table 1: Confusion Matrix and Error Rate of KNN Classification Splits

If we are to look at the values shown in Table 1, we notice that the 60/40 split has the lowest error rate of 6.9% among the three splits whereas the 80/20 split has the greatest error rate of 20%. If we are to look at the F1-Scores of the 60/40 split for both individuals with and without cancer, it is in the high 90 and almost full for those with cancer, hence indicating an accurate model. This is evident throughout all three models which have extremely high precision rates.

Decision Tree Classification

The second modelling technique utilised in this project is the Decision Tree classification. This classification is again split into three distinct test/training splits which help us determine the precision by which our model functions.

Decision Tree Split Accuracies and their Confusion Matrix

Table 2 showcases the confusion matrix and error rates of all the splits done during Decision Tree Classification. To understand the splits, the value on the left of the '/' indicates the training set split and the one on the right represents the testing set split.

	80/20 Split	60/40 Split	50/50 Split
Confusion Matrix	[[6 4] [1 4]]	[[14 3] [9 3]]	[[18 3] [8 7]]
Error Rate	33.33333333333333 %	41.37931034482759 %	30.555555555555557 %
F1-Score for Individuals Without Cancer	0.71	0.74	0.77
F1-Score for those with Cancer	0.62	0.50	0.56

Table 2: Confusion Matrix and Error Rate of Decision Tree Classification Splits

If we are to look at the values shown in Table 2, we notice that the 50/50 split has the lowest error rate of 30.6% among the three splits whereas the 60/40 split has the greatest error rate of 41.4%. These are significant error rates.

Feature Selection

Feature selection with the help of Hill Climbing provided us with a good subset of features. This provided a fairly negligible difference in accuracy compared to the difference with all the variables taken into consideration. Removal of redundant features did not do much to the accuracy of the model, mainly because it was extremely accurate from the get-go.

Conclusion

Although the precision is high for both classification methods, the K-Nearest Neighbour classification surpasses the Decision Tree classification in both precision and error rates significantly. Therefore, the most apt classification model for this dataset would be the KNN classification out of these models. There is scope for improvement for the accuracies for this dataset, but it is out of scope of this course. Algorithms such as the Random Forest Algorithm are extremely accurate but do tend to have a higher run time and is more complex.

References

Cervical Cancer Data Set. Extracted from https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk

Metrics SciKitLearn API. Confusion matrix, K Graph. Extracted from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Nearest-Neighbour Classifier SciKitLearn API. Extracted from https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html

K-Nearest-Neighbour Classifier SciKitLearn API. Extracted from https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier

Decision-Tree Classifier SciKitLearn API. Extracted from https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier

Appendix

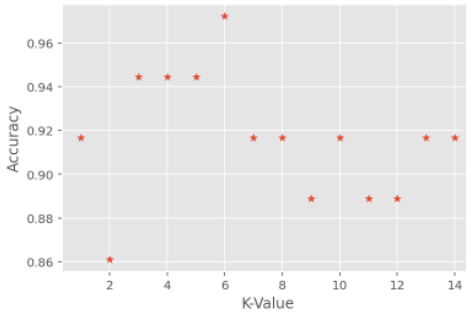


Figure 3.2.1: Plotting the K-Values for 50/50 Split

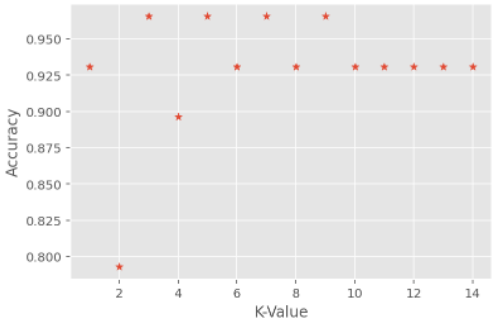


Figure 3.2.2: Plotting the K-Values for 60/40 Split

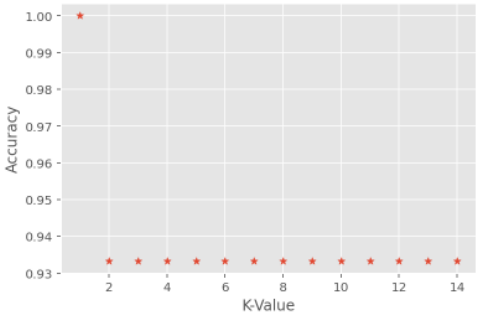


Figure 3.2.3: Plotting the K-Values for 80/20 Split